Frontiers in Computer Vision and Pattern Recognition (2023)
Clausius Scientific Press, Canada
DOI: 10.23977/fcvpr.2023.010102
ISSN 2522-6665 Vol. 1 Num. 1

Research on Smart Site Safety Hazard Warning Technology Based on Yolov5

Yuanyuan Wang[1,2,*], Jiahui Cao[1], Mingran Qi[1], Sisi Liu[1], Xiuchuan Chen[1]

[1]School of Computer and Software Engineering, Huaiyin Institute of Technology, Huai'an, 223001, China
[2]Jiangsu Provincial Engineering Laboratory of Mobile Interconnection Technology for Internet of Things, Huai'an, 223001, China
[*]Corresponding author

Keywords: Smart site, behaviour recognition, deep learning, Yolov5, PyQt5

Abstract: Smart Site is an emerging construction industry project management concept, whose main feature is the integration of modern information technology means to help construction safety. The construction industry is a safety accident-prone industry, and due to its complexity and specificity in the production process, it is difficult to truly conduct real-time inspection of safety hazards at construction sites by relying solely on supervisory and managerial safety officers and manual inspections.Aiming at the safety of construction site workers nowadays, in order to reduce potential hazards and considering the poor adaptability and high false detection rate of current fall detection methods, a technical study of fall detection at smart construction sites based on target detection is proposed. This study uses a home-made web public image dataset, the runtime window is optimised using PyQt5 GUI, the algorithm framework uses Yolo v5 training model, and the trained model achieves the recognition of different scenes and different poses of falls, providing a certain code theoretical basis and feasible model support for the development of a system related to abnormal behaviour detection of construction site workers. The experimental results show that the method proposed in this experiment has high accuracy in human fall detection, providing a certain code theoretical basis and feasibility model support for the development of the system related to abnormal behaviour detection of construction site workers in terms of accuracy of measurement, which is more suitable for construction units.

1. Introduction

As the scale of construction sites in China continues to expand, the safety of workers is beginning to be of increasing concern. To date, there has been a relatively mature development of computer vision in various places and a long accumulation of human behaviour recognition.

Zhang et al.[1] has proposed a scheme in a past study where the first step was to construct a spatio-temporal attention module based on non-local operations. The addition of joint and skeletal information to the spatio-temporal attention module optimises the model and provides a feasible example reference and thought guide for the implementation of this design.

In the article "Research on Indoor Dangerous Behaviour Monitoring and Early Warning System for the Elderly"[2,3], key points of the skeleton are extracted for common dangerous behaviours such as abnormal fall behaviour while RGB cameras collect posture information. Support vector machines are used to identify and classify various types of abnormal human behaviour, which are then combined with trajectory and location information to provide remote warning and emergency calls to relevant personnel. Reference examples are provided for the design of related systems implemented by this method.

Computer vision has been used for some time and practice to recognise human actions, there are familiar ones based on WiFi signal perception[4], infrared perception, etc. However, because of the fast speed and high similarity of human actions, current algorithms cannot obtain high recognition rates. In 2019, Xiaoye et al.[5] proposed a new method for this purpose based on multi-scale directed depth motion maps (MsdDMMs) and Log-Gabor filters for a new method of human action recognition. The method proposes MsdDMMs in an energy framework depending on the speed and temporal order of the actions. In the same year, Wang's [6] team summarised WiFi CSI (channel state information) based behaviour recognition[7]. pengcheng et al.[8]also proposed a 3D-CNN based IC3D for the problem of low accuracy of human behaviour in video neural network in human behavior recognition[9], which regularizes the whole network to prevent overfitting and reduces one fully connected layer layer.

In 2020, Jia et al.[7] proposed dual-stream temporal convolutional networks (TSTCNs) [4] that make full use of inter-frame vector features and intra-frame vector features of skeletal sequences in spatio-temporal representations, a framework that integrates different feature representations of skeletal sequences so that the two feature representations can compensate for each other. The team of Ushapreethi[9], who found a for an unmanned environment successful human action recognition (HAR) system[10], proposed three key steps for the improvement of HAR systems to improve the accuracy of existing HAR systems. A method based on a time-range Doppler point network[4,11] is proposed to analyse human behaviour in the paper Human behaviour recognition based on distance velocity and time points, where the time point network can learn structural features from micro-motion trajectories more efficiently than by directly processing the raw point cloud.

Current deep learning-based methods have demonstrated impressive performance, but the trade-off between efficiency, robustness and accuracy in existing methods remains inevitable. In 2019, a novel bidirectional optimally coupled ligh-tweight network for efficient and robust multi-person 2d pose estimation is presented in the paper Bidirectional Optimally Coupled Lightweight Network(BOCLN)[10]architecture that gives a probabilistic limb heat map to represent the spatial context of the input joints and guide the overall pose skeleton prediction, enabling the best possible accuracy and robustness of pose estimation for each person in a cluttered scene (involving a crowd).In 2020, to improve the accuracy and robustness of human pose estimation, Liang et al. proposed a multi-residual modular stacked hourglass network[12](MRSH) Soon after, Li's team combined the affinity module and residual attention module into a stacked hourglass network [13,14]for human pose estimation, with more accurate and robust human pose estimation results in images with complex backgrounds. To achieve sufficient recognition accuracy, a novel Yolo v4+LSTM [15,16]network[17]was proposed by Nguyen et al. In 2021, the article[18]on efficient estimation of human pose using parallel pyramid networks proposed the design of a "parallel pyramid " network (PPNet), in order to have a better balance between operational speed and accuracy. The quest for higher accuracy opens the research of this paper.

2. Fall Detection

2.1. General Structure

The research implementation of the worker abnormal behaviour detection method in the face of smart construction sites is based on three main modules, the data set, the model building and training and the code running interface PyQt optimization. The main flowchart for detection is shown in Figure 1: we can pass a video or photo of the person through the Yolo v5m model and subsequently the fall can be indicated by the interface display.

Figure 1: Flow chart of fall detection

2.2. Yolo V5 Model

YOLOv5 is the latest network in the YOLO family of algorithms, consisting of an input, a feature extraction network, a special fusion network and an output. The yolov5m structure is used to meet the relevant requirements, achieve the expected model training effect, and improve the accuracy of the model in discriminating human body. YOLOv5 continues the multi-scale feature fusion approach of FPN+PAN in YOLOv4, with three scales of detection layers. Compared to YOLOv4, YOLOv5 has a slightly reduced detection accuracy, but a greatly improved training time and detection speed, and is much more flexible. This paper is oriented towards the training of human abnormal behaviour recognition using the Yolo v5 model budget method for the study of worker fall detection methods for smart construction sites.

(1) Model input port

The input port features Mosatic data enhancement, adaptive anchor box border calculation and adaptive image sizing and scaling. Mosaic data enhancement and adaptive anchor box border calculation are performed on the input image.

Mosaic data enhancement.

The mechanism for data enhancement is random free cropping, random scaling, random permutation and stitching of the input image, which can improve the success rate of target detection.

Adaptive anchor box border calculation.

When performing training, it is a critical step that each training can be autonomously adapted to calculate the best anchor box value for the target based on the different training sets used for training. In yolov3 and yolov4, training different data sets are run through a separate procedure to obtain the initial aiming frame. However, in yolov5 this has been embedded in the code.

(2) Backbone

The backbone is the backbone of the network, a convolutional neural network that extracts features from images and aggregates them to form a variety of fine-grained image features, including focus, conv, bottleneckcsp, spp. The Conv module is composed of conv, Bn, and Leaky_relu activation function. bottleneckcsp is composed of bottleneck, CSP1, and bottleneck residual structure bottleneck, which contains a 1×1 convolution, which reduces the number of parameters and therefore the computational effort of the model. Spp, on the other hand, contains three components: conv, maxpooling, and connat.

The Spp module fuses 512 of 13×13 inputs according to a maximum pooling of 13×13, 9×9, and 5×5 for multi-scale eigenvalue fusion operations.

(3) Neck

Neck uses the FPN+PAN construction and is the intermingled part of the network. Neck mixes and combines the input features and then feeds them into the Precision layer. The FPN structure facilitates the optimisation of the lower layer feature propagation due to its top-to-bottom transmission of strong features. The feature pyramid, on the other hand, consists of 2 PAN structures, which are characterised by their bottom-up transmission properties. The combination of FPN and feature pyramid operation together with the addition of CSP2 is very beneficial for the enhancement of the network feature fusion capability.

(4) Prediction

The structure of Prediction is shown in Figure 2.

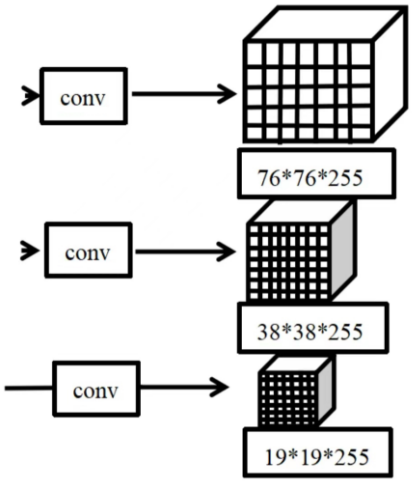

Figure 2: Structure diagram of prediction

The function of the target detection task is generally composed of two parts: the classification loss function and the regression loss function. In this paper, CIOULoss is used as the loss function of the Bounding box in yolov5, so that the speed and accuracy of the prediction box regression can be higher and the accuracy we obtain can be improved.

2.3. CIOU Loss Function

A good target frame regression function has these three key geometric factors: overlap area, centroid distance, and aspect ratio. The closer the aspect ratio is to the target box, the better that prediction box will be. As well as the two issues of how to make the regression more accurate when the prediction frame and target frame overlap, the algorithm equation (1) is given as:

$$CIOU_{L}\text{oss} = 1 - CIOU = 1\text{-}(IOU - \frac{D\text{istance}2^2}{Dis\tan ceC^2} - \frac{v^2}{(1-IOU)+v}) \tag{1}$$

where v is a parameter that measures the consistency of the aspect ratio, defined as:

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{W^{p}}{h^{p}})^2 \tag{2}$$

The three items of the CIOU correspond exactly to the calculation of the IOU, the distance from the centre point and the aspect ratio.

2.4. DIOU Nms

In the classical NMS algorithm, IOU is the only factor considered. However, in practice, errors of missed detection often occur, when two different objects are close to each other, which, due to the large IOU value, can lead to only one detection box at the end after processing by NMS.

DIOU_nms also takes into account the distance between the centroids of the two frames when the model detects inputs with overlapping targets, because of the overlapping obscured targets. DIOU_nms is superior to the classical NMS when dealing with obscured targets. If the IOU between two frames is large and the distance between their centres is also large, they may be considered as frames of two objects and therefore not filtered out.

3. Analysis of Experimental Results

3.1. Experimental Environment and Data Set

The algorithms in this paper use python as the development language. Training and testing were performed on a window11 system with a hardware environment of Intel(R) Core(TM) i5-10210U CPU@1.60GHz 2.11 GHz, NVIDIA GeForce MX250, and a graphics card with 6G of shared memory, which was able to meet the minimum requirements for deep learning.

The experiments were conducted using a public dataset containing 1442 original fall images and the corresponding labeled images can be used to perform model training. Before doing the experiments, the images were processed as necessary. The focus module first copied four copies of the input and sliced them into four slices, the concat layer stitched the images together while increasing the number of features in the images, keeping the information under each feature unchanged, and finally the feature images were sliced and the original three 4×4 images were split into twelve 2×2 feature images by the slicing operation.

3.2. Algorithm Evaluation Criteria

For the experimental training and detection effect in the fall scene experiment, this paper cites Accuracy as the evaluation index, using the model in the dataset to verify Accuracy is the accuracyof the overall judgment of the classification model. Generally speaking Accuracy is proportional to classifier performance, i.e. the higher the Accuracy, the better the classifier performance, which is calculated as follows (3):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Overall Accuracy (ACC), it is very incorrect and comprehensive to evaluate the performance of a model classification based on ACC alone.

Precision is analysed from the point of view of how much of the data with a positive predicted outcome has been correctly predicted, calculated as follows (4):

$$precision = \frac{TP}{TP + FP} \tag{4}$$

Sensitivity or recall is a measure of how many of the data in the test set that are positive classes have been correctly predicted, which measures the ability of the classifier to discriminate between positive classes.

$$Sensiticity = \text{Recall} = \frac{TP}{TP + TN} \tag{5}$$

Specificity is the proportion of all true negative examples that are predicted by the classifier to be negative classes, which measures the ability of the classifier to recognize negative examples(Sun et al.2020).

$$Sensitivity = \frac{TN}{TN + FP} \tag{6}$$

The F1-score (F1-score) has not only Precision but also Recall and therefore turns out to be one of the measures of classification problems. It is the summed average of Precision and Recall for that category, and has a value between 0 and 1. For the behavioural categories of this model, the higher the F1-score is usually the better the model performance, considering both its accuracy and recall, and it is calculated as follows:

$$F_1 Score = \frac{2 * \text{Pr}ecision * \text{Re}call}{\text{Pr}ecision + \text{Re}call} \tag{7}$$

4. Experimental Results

There is a positive correlation between confidence and accuracy. If a higher accuracy is desired, a higher confidence can be set. However, when the confidence is set too high, the detection effect for multiple target inputs will be poor, resulting in many targets not being displayed in the detection result box due to not reaching the corresponding accuracy. When the confidence is too low, there may be a large number of inaccurate predicted targets for multiple target inputs, resulting in many inaccurate target objects, thus lowering the overall prediction effect of the model. Therefore, the confidence setting should be dynamically adjusted according to the actual input situation and the desired effect as show in Figure 3.

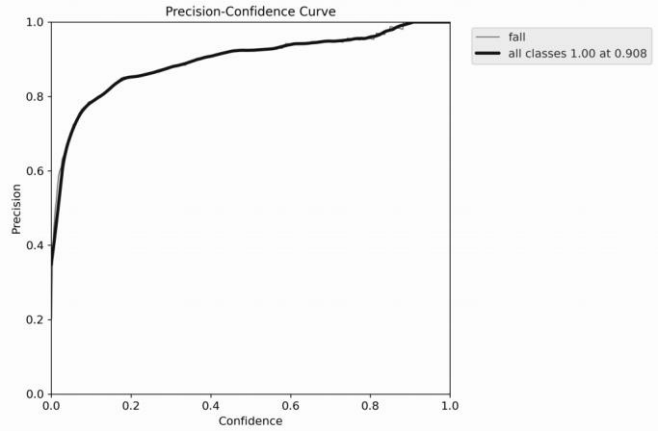

Figure 3: Plot of confidence versus precision

As shown in Figure 4, at a confidence level of 0.176, the model achieves a peak F1 score of 0.84 for all categories.

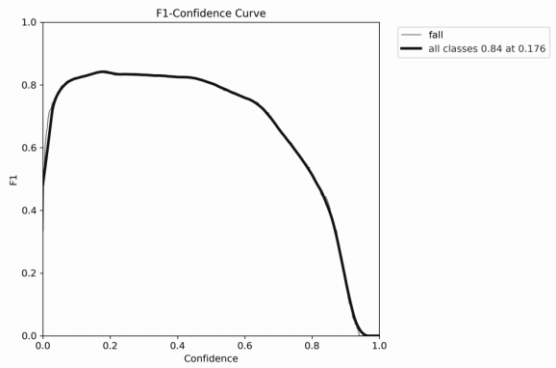

Figure 4: Plot of model confidence versus F1 score

In order to verify the robustness of the algorithm model in terms of fall detection, Yolov5 was selected for experimental comparison with Yolov5's on the corresponding dataset with the improved network in this paper, and the experimental comparison results are presented in Table 1 for the experimental accuracy and F-socre at the same IoU of 0.5 threshold, respectively.

Table 1: Experimental comparison of different Yolov5s

Model	Accuracy	Recall rate	f-score
Yolov 5	42.1%	73%	53.4%
This article	45%	82%	58.1%

The traditional Yolov5 had an accuracy of 42.1%, a recall of 73%, and an f-score of 53.4%, while the improved accuracy of 45%, recall of 82%, and f-score of 58.1% were all much higher.

5. Conclusion

This methodological study produces its own dataset, selecting a multi-angle, all-round and multi-sample feature map of fall behaviour samples from different sources such as public datasets, web images and excerpted web video screenshots. To improve the accuracy of the model, two training sessions were conducted to optimise the model based on the first overfitting model, reducing the error of model overfitting on the accuracy of human behaviour recognition. The model achieves better accuracy for different inputs, with good overall model performance and a smoother interface running under CPU with camera input, achieving the expected real-time detection effect. To further improve the results, the next research direction may be to increase the sample complexity and number, or to improve the model.

Acknowledgement

This work was supported by Huai'an City Science and Technology Plan Project (grant no. HABL202101), Huaiyin Institute of Technology - Huai'an Economic and Technological Development Zone Industry-University Research Cooperation Project (grant no. Z413H21522), Innovation and Entrepreneurship Training Programme for University Students (grant no. 202211049087Y, 202211049268XJ), Enterprise collaboration cross-project (grant no. Z421A22349, Z421A22304, Z421A210045).

References

[1] Zhang Jiaxiang, Liu Ruhao, Jin Chenxi, and so on. Skeleton behavior recognition by combining spatiotemporal attention mechanisms and adaptive graph convolutional networks. Signal processing, 2021, 37 (7): 1226-1234.
[2] Wu Haoyuan, Xiong Xin, Min Weidong, Zhao Haoyu, Wang Wenxiang. Behavioral recognition method based on

multilevel feature fusion and time domain extension. Computer Engineering and application: 1-10 [2022-02-13].
[3] Wang Lingling, Guo Shiqi, Zhou Ying, Chen Kunhui. Study on the indoor risk behavior monitoring and early warning system for the elderly. Information technology of civil and construction engineering: 1-9 [2022-01-22].
[4] Li M, Chen T, Du H. Human behavior recognition using range-velocity-time points IEEE Access, 2020, 8: 37914-37925.
[5] Xiaoye Zhao, Xunsheng Ji, Yuanxiang Li, Li Peng. Combining multi-scale directed depth motion maps and log-gabor filters for human action recognition. Journal of Harbin Institute of Technology (New Series), 2019, 26 (04): 89-96.
[6] Wang Z, Jiang K, Hou Y, et al. A survey on human behavior recognition using channel state information. Ieee Access, 2019, 7: 155986-156024.
[7] Jia J G, Zhou Y F, Hao X W, Li F. Two-stream temporal convolutional networks for skeleton-based human action recognition. Journal of Computer Science and Technology, 2020, 35 (3): 538-550.
[8] Pengcheng D, Siyuan C, Zhenyu Z, Zhigang Z, Jingqi M, Huan L. Human behavior recognition based on IC3D//2019 Chinese Control And Decision Conference (CCDC). IEEE, 2019: 3333-3337.
[9] Ushapreethi P, GG L P. Skeleton-based STIP feature and discriminant sparse coding for human action recognition. International Journal of Intelligent Unmanned Systems, 2020.
[10] Li S, Fang Z, Song W F, Hao A M, Qin H. Bidirectional optimization coupled lightweight networks for efficient and robust multi-person 2D pose estimation. Journal of Computer Science and Technology, 2019, 34 (3): 522-536.
[11] Bao W, Yang Y, Liang D, Zhu M. Multi-residual module stacked hourglass networks for human pose estimation. Journal of Beijing Institute of Technology, 2020, 29 (1): 110-119.
[12] Chen Lumeng, Cao Yan Yan, Huang Min, Xie Xingang. Flame detection method based on an improved YOLOv5. Computer Engineering: 1-17. 2020.
[13] Fan Wenshuo. Research and Design of fixed point detection technology based on image recognition. University of Electronic Science and Technology, 2021.
[14] Hua G, Li L, Liu S. Multipath affinage stacked—hourglass networks for human pose estimation. Frontiers of Computer Science, 2020, 14 (4): 1-12.
[15] Lu J, Nguyen M, Yan W Q. Deep learning methods for human behavior recognition//2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). IEEE, 2020: 1-6.
[16] Lu Yong, Lu Zhaohe, Wang Xiaodong, Zhou Xingming. Review of research on human behavior perception techniques based on WiFi signals. Journal of Computer Science, 2019, 2: 231-251
[17] Sun Bo, Yang Lei, Guo Xiumei, Chen Ran, Zhang Tong, Jia Hao. ECG signal identification method based on hybrid CNN and SVM. Journal of Shandong Agricultural University (Natural Science Edition), 2020, 51 (02): 283-288.
[18] Zhao L, Wang N, Gong C, Yang J, Gao X. Estimating human pose efficiently by parallel pyramid networks. IEEE Transactions on Image Processing, 2021, 30: 6785-6800.